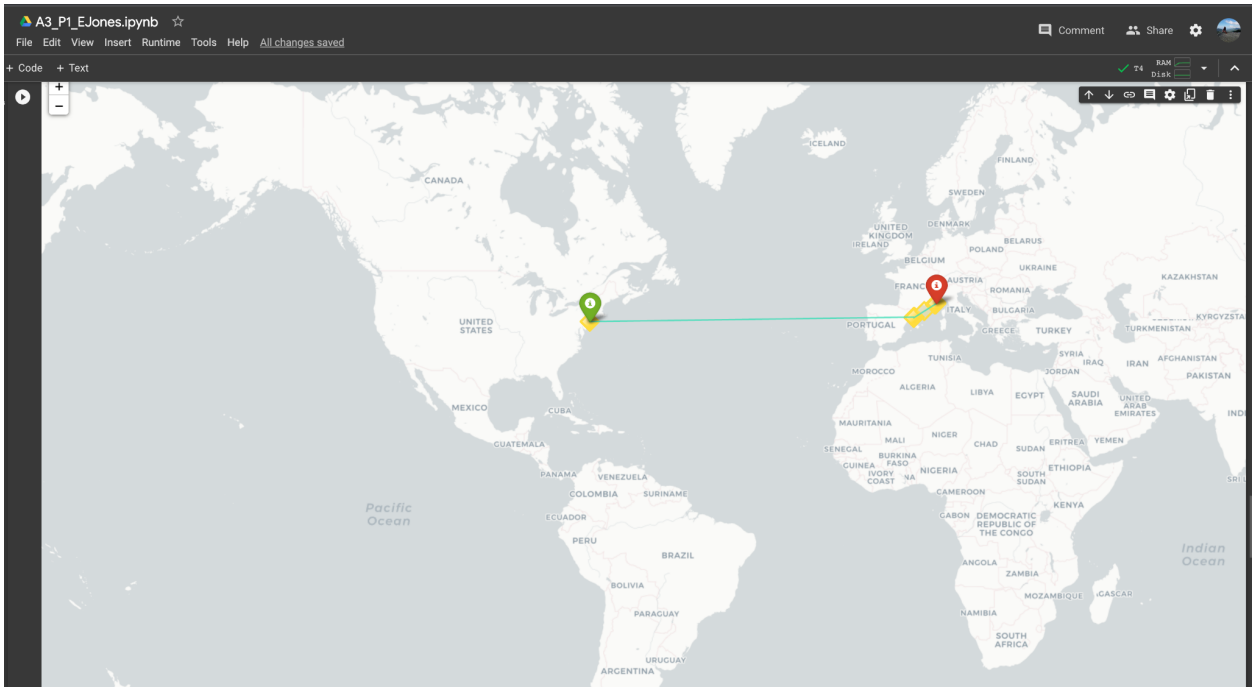
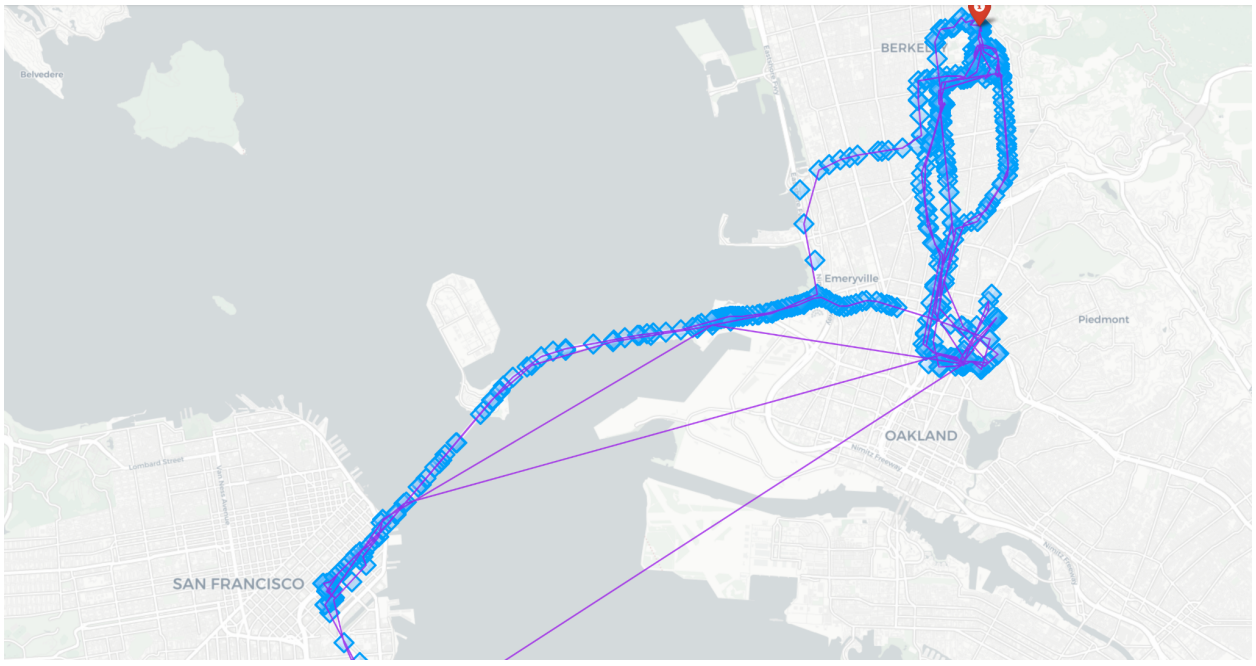
Erin Jones - Assignment 3 Part 1 - Visualizations

06/01 - 06/10/2022



I was still in Washington, DC at that time and then I flew to Barcelona for a music festival and then on to the south of France to meet my mom and spend some time in the Cote D'Azur. So much more exciting than my last week - see below - I guess I am trying to shake off some mid-term stress via nostalgia for stress-free times.

10/1 - 10/7/2023



My life has shrunk to exist in Berkeley, Oakland, and San Francisco. So Sad 😕

Access the Collab notebook I used via this link – https://colab.research.google.com/drive/1VL-yzP5p8AnpkWQ9pZFhbrZ9MsmzDqUI?usp=sharing

Project Preparation (50 pts)

Team Members: Balaji Balaganesan , Breitling Snyder , Erin Engle Jones
Deadline: 10/13/2023 (10/15 Deadline Extension Approved)

Part I: Search and Selection - Motivate your research project answering the three questions below **(30 pts)**

- **What is the problem?**

  Determining the impacts of transportation-related policy on a given location's infrastructure, carbon footprint, and populace behaviors is difficult when examining it using a framework whereby a policy is chosen and then data is turned to for the purposes of mining its effects. The implementation of a policy could be compared to a field or a natural experiment; very few (if any) of the extraneous variables can be controlled and thus it is very difficult to derive a causal relationship between a policy and any particular outcome. There are any number of confounding variables that could affect said outcome, from overlapping policies to the methods and timing of implementation to cultural aspects of the region. Additional confounding questions include - What is the radius of impact for a particular policy? What is the appropriate time period after which we can consider the experimental treatment (i.e. policy implementation) to be complete, or to see effects? The latter is especially applicable when considering investments in infrastructure that sometimes take decades to build.

  National policies and infrastructure often have very large samples to use for reference, but much policy related to transportation and infrastructure is implemented in a far more localized fashion. Furthermore, local, county and state-level budgets are much smaller than those of the federal government and thus the concern over the impact of each dollar becomes more acute. Staff capacity to evaluate, design, and/or implement policy may also be more severely limited. Case studies are often used as blueprints for local and regional policy design and implementation. However, case studies can be difficult to identify or apply in new regions with the same outcomes for a number of reasons, including differences in existing infrastructure, as well as the cultural and demographic characteristics in populations. Evaluation of case study policy outcomes can also be limited if administrative capacity is lacking. It can be difficult for cities and/or regions to identify an "optimal" case study, and adjust it accordingly to fit their locale's needs and meet planned objectives.

- **Why should we care?**

In California alone, transportation-related emissions make up over half of the State's total polluting greenhouse gas emissions. Throughout the U.S., it tends to be one of the largest sectors of our total carbon footprint.

In light of the climate crisis and its politicization, efforts to control carbon emissions and promote sustainable human behavior through transportation mode shift are both of grave importance in achieving IPCC goals and under significant political scrutiny (i.e. was a policy the most appropriate use of budget?) (Shapiro et. al). The need for effective policy is paramount. But how does one identify what is an "effective" policy?

Data on carbon emissions, transportation infrastructure, and human behavior related to transportation infrastructure exists. However, applying machine learning to examine their impact is a relatively nascent approach due to the recent advancements making things like unsupervised clustering algorithms easy and inexpensive to implement. We aim to take such an approach, and hopefully uncover new insights on how to identify efficacious policies, and how existing infrastructure and population demographics shape the implementation of such policies.

- **What do you want to do applying data analysis and modeling? + Part II: Data Science Story (10 pts)**

  We aim to use our data analysis to identify changes in a location's emissions and/or commute behaviors over time, and see if that change can be linked to any relevant policies or action taken by public agencies in said location. We then will identify if there are any particular characteristics about a location that make policy implementation successful, such as transportation network structure or population demographics. We identified the following papers to guide our areas of concern and methods of data analysis.

  Cited Papers:

  - Boeing, Geoff. "OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks." Computers, Environment & Urban Systems, vol. 65, 2017, pp 126-39. ScienceDirect, https://doi.org/10.1016/j.compenvurbsys.2017.05.004
  - Cutler, Adele, and Leo Breiman. "Archetypal Analysis." Technometrics, vol. 36, no. 4, 1994, pp. 338–47. JSTOR, https://doi.org/10.2307/1269949
  - Jones, Christopher and Kammen, Daniel. "Spatial Distribution of US Household Carbon Footprints Reveals Suburbanization Undermines Greenhouse Gas Benefits of Urban Population Density." Environmental Science & Technology.

vol. 48, no. 10, 2013, pp 895-902. ResearchGate, https://doi.org/10.1021/es4034364

- ServiceInnovation Lab. (2023b). *Machine learning introduction - Smart Policy Analysis Tools*. Smart Policy Analysis Tools. https://serviceinnovationlab.github.io/bagel-box/intro/
- Shapiro, H. T., Diab, R., Brito Cruz, C. H. de, Cropper, M., Fang, J., Fresco, L. O., Manabe, S., Mehta, G., Molina, M., Williams, P., Zakri, A. H., & Winnacker, E.-L. (2010). (rep.). *Climate Change Assessments | Review of the processes and procedures of the IPCC* (pp. 1–123). Amsterdam, ND: InterAcademy Council. https://archive.ipcc.ch/pdf/IAC_report/IAC%20Report.pdf
- Tao, Tao, and Jason Cao. "Using Machine-Learning Models to Understand Nonlinear Relationships Between Land Use and Travel." Transportation Research Part D: Transport and Environment. vol. 123, 2023. ScienceDirect, https://doi.org/10.1016/j.trd.2023.103930

The state of affairs for infrastructure, behavioral patterns, demographics, and carbon emissions for a particular location at a given time can be represented as an n-dimensional vector, whereby each element is a quantified representation of some characteristic of that location. For example, Oakland CA could be represented as a vector with the following features:

- Network characteristics based on transportation infrastructure at that time (node and edge attributes, centrality, clustering and path length attributes). These will be identified using OSMnx.
- General population demographics, infrastructure spending data and density attributes, from ACS.
- Human behavior in interaction with infrastructure (commute behavior, etc), from ACS data.
- Carbon emissions data for the given time period, from the DARTE/CoolClimate dataset.

If one were to create vectors representing block groups or neighborhoods within a county and create vectors for various years, these vectors could then be fed into a clustering analysis (e.g. k-means, hierarchical or archetypal analysis) to determine if the dimension of time seems to bear any impact on the clusters which form. If there is a divergence that seemingly coincides with movement from one year to another, we will further examine if any transportation related policies were implemented between the given years.

We also plan to complete an analysis across policy-making areas, using the same characteristic vectors resampled at the county granularity. We will then examine clusters with the lowest levels of CO2 emissions so see if the selected locations implemented

particular types of transportation policies in the given year or the 5 years prior. We will also analyze whether other factors may have contributed to the change, in addition to or rather than policy implementation. For location vectors where a change has been identified, we can use further clustering techniques to see how the existing transportation network structure and population demographics may have an impact on the effectiveness of a policy intervention.

The more time consuming part of this analysis will come from engineering features, normalizing vectors appropriately for a given analysis, and then manually exploring policy decisions for chosen vectors.

- **Identify your data sources + Part III: Data Gathering (10 pts) - Find a data set to work with it and describe it here**.

Dataset	Link	Description
ACS 5-year estimates	https://www.census.gov/topics/employment/commuting/guidance/commuting.html; https://data.census.gov/table	American Community Survey (part of the Census data) – provides estimates based on survey data about commute behavior, population demographics, and more for all geographies throughout the U.S. and Puerto Rico.
DARTE	https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1735	The basis for the CoolClimate Dataset (Dan Kammen). Ideally we will be granted access to the CoolClimate Data, but in case we aren't we will use the DARTE dataset that contains details on carbon emissions for road transportation
OpenStreetMap + OSMnx	https://osmnx.readthedocs.io/en/stable/	OpenStreetMap is an open source collaborative geospatial dataset, maintained by volunteers. OSMnx is a Python package developed to

		access and analyze transportation networks (eg. streets, bike lanes, and transit networks) data from OpenStreetMap

Note: We will be pulling policy data from various sources for a sample of locations based on clustering outcomes. We also may deem it to be necessary to add additional features from additional sources (and are open to your suggestions). The data science process requires flexibility when it comes to raw data.